图书情报工作
第64卷 第20期 2020年10月

人工智能领域科研团队识别与领军团队提取*

余厚强 白宽 邹本涛 王曰芬
南京理工大学经济管理学院 南京 210094

摘 要:[目的/意义]对人工智能领域科研团队进行识别,并基于多个维度的指标提取领军科研团队,旨在丰富科研团队识别的流程与方法,为从科研团队视角分析人工智能领域脉络、前沿和主题提供依据。[方法/过程]以 Web of Science 为数据来源,采集 2009 - 2018 年间人工智能学科领域所有科技论文的数据,通过算法设计与人工核查进行数据清洗;基于分数计数法构建全局合著网络,并利用社区探测算法动态调参、识别科研团队;进而基于多维度的指标提取出领军团队,并加以比较分析。[结果/结论]从实践出发构造人工智能科技论文数据清洗的规则;构建基于合著关系识别人工智能科研团队的流程体系;提出通过消除边缘结点进行合著网络筛选,进而利用已知团队作为参考进行参数调整的思路;较为系统和准确地识别出全球人工智能科研团队,并基于发文量、被引量、h 指数、中介中心度、接近中心度和加权点度中心度 6 个维度的指标提取出领军科研团队,同时,给出结合论文数据和实证调研对每个领军团队的示例性分析。

关键词:人工智能 合著网络 科研团队 领军团队 数据分析
分类号:G250
DOI:10.13266/j.issn.0252-3116.2020.20.001

1 引言

在大科学时代,科研合作被视作提高科研效率的重要手段。科研合作的一种主要表现形式是科研团队的形成,它是科学共同体的重要构成。科研团队不仅是科学研究的中坚力量,能够体现出一个学科领域人力投入的集聚程度,而且引领着科学研究发展的态势与前沿。因此,科研发展特点与规律的研究中需要关注科研团队,科技政策制定与调整中更需要密切关注科研团队及其所做的研究。此外,在科技评价中,尤其是基金支持、高校人才引进等方面,也往往需要结合科研团队进行考核与评价。

科研团队是指"以科学技术研究与开发为内容,由优势互补、愿意为共同的科研目的、科研目标和工作方法而相互协作配合承担责任的科研人员组成的群体"[1]。本文所研究的对象是建立在科研协作关系上的虚拟科研团队。传统识别科研团队的方法主要是专家访谈、问卷调研等,随着社会网络分析方法的发展,利用合著网络关系识别科研团队得到深入研究。科研团队识别的相关研究主要从两个方面展开:

(1)识别不同领域的科研团队。由于科研团队在现代科学研究中的重要性,不同学科领域均关注科研团队问题,例如管理科学[2]、流行病学[3]、情报学[4]、肿瘤学[5]等学科,均研究了科研团队识别的问题。

(2)改进科研团队识别的算法。最基本的识别算法是社会网络分析中的派系识别方法,在后来的发展过程中,有运用向量空间模型[6]、引入关联规则挖掘中 FP-Growth 算法[7]、基于原始数据矩阵因子分析[8]、合著网络加权[9]等方法开展科研团队识别的实证研究。

本研究主要有两个出发点:①人工智能领域的重要性日益增强,但是还没有专门针对该领域科研团队的系统性研究。人工智能技术的飞快发展,在全球范围内引起广泛重视,不仅集聚了许多学科领域学者参与到研究中,而且随着各个国家和地区对于人工智能产业的大力布局,人工智能的研究也不断深化与发展。因此,从科研团队角度对人工智能领域展开分析具有重要决策支持意义。②既有的科研团队识别研究主要是基于较小数据规模的实证研究,鲜有基于大规模数

* 本文系国家社会科学基金重大项目"面向知识创新服务的数据科学理论与方法研究"(项目编号 16ZAD224)研究成果之一。
作者简介:余厚强(ORCID:0000-0002-9241-6630),副教授,博士;白宽(ORCID:0000-0003-4048-6941),硕士研究生;邹本涛(ORCID:0000-0002-3972-0705),博士研究生;王曰芬(ORCID:0000-0002-7143-7766),教授,博士生导师,通讯作者,E-mail:yfwang@njust.edu.cn。
收稿日期:2020-05-11 修回日期:2020-06-24 本文起止页码:4-13 本文责任编辑:杜杏叶

据的实证分析。已有的研究主要选取数量有限的期刊作为数据来源,识别出的科研团队数量在几个到几十个不等。并且,采用的方法主要是先确定团队领导者,再扩展得到团队成员,这样虽然避免了团队规模无法确定的难题,但是少数极大团队的存在会导致按中心度排名,许多其他团队无法得到显示。

本文旨在解决以下主要研究问题:

(1)如何基于大规模科技文献数据识别出人工智能领域科研团队?要回答这个问题,不仅需要解决机构数据、作者数据的规模化清洗问题,而且需要通过试验确定合适的科研团队分割粒度。

(2)基于识别出的大量科研团队,从不同角度提取出的人工智能领域领军团队有哪些?具体来说,将选取不同的指标,从不同角度去提取人工智能领域的领军团队。

领军科研团队是指在科学研究上成绩突出的科研团队,简称为领军团队。科研团队的评估是复杂的,不宜对多样化的科研团队使用一种指标或通过加权降维去排名展示。因此,本文认为不同维度表现突出的科研团队都是领军团队,这既可以是单一维度上表现突出,也可以是若干维度上同时表现突出。在系统识别出科研团队后,从多维视角提取出领军科研团队,对于分析全球人工智能的研究力量及其分布、跟踪研究发展状态与前沿、制定研发规划等,具有重要的决策支持意义。

2 研究设计

2.1 整体流程设计

从大规模数据集中识别出科研团队的过程,实际上就是数据采集、处理、挖掘与利用的数据分析过程。因此,本研究从数据驱动出发,将涉及到的主要内容嵌入到操作环节中,设计科研团队识别的整体流程,如图1所示:

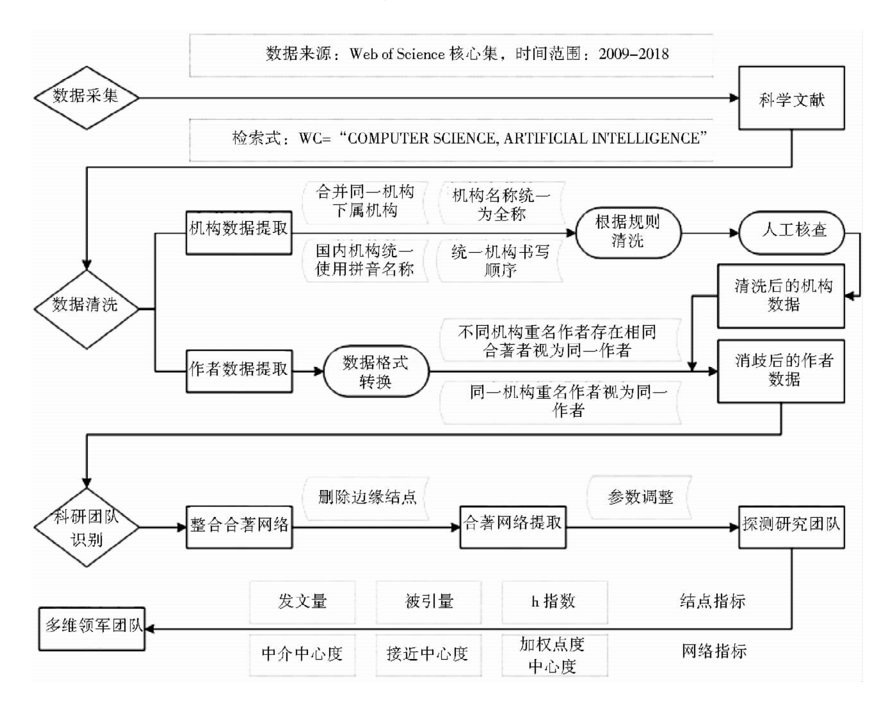


图1 基于数据分析的人工智能科研团队识别流程

2.2 数据采集

本研究采用Web of Science(WoS)的数据进行分析,WoS是国内外进行科学计量分析最权威的数据库之一。由于索引的期刊经过专家遴选,通常被认为是领域核心期刊,具有较高的质量。并且,大多数WoS索引的期刊具有较好的国际导向,便于分析和比较全球范围的科技文献发表情况。因此,采用WoS数据具备较好的可靠性和可信度。

人工智能是个复杂的新兴领域,为了检索获得该主题的所有相关文献,采用关键词检索将遇到检全率不足的问题,因为人工智能涉及非常多的子主题。但是一旦选用了过多的关键词,由于不少关键词有歧义或范围太广,又会出现检准率的问题。而WoS数据库的学科分类中在计算机大类下设有人工智能子类,涵盖了所有与人工智能密切相关的期刊,因此,我们将该子类的所有文献下载下来。虽然WoS数据库的学科

分类是基于期刊的，存在局限性，但该学科分类采用专家同行评议实现，具备较好的可信度。经过比较，利用该学科分类进行检索的效果最好。因此，采用“WC = ‘COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE’”检索式进行检索，检索时间范围为 2009 – 2018，检索时间是 2019 年 1 月 16 日，共采集到 421 148 篇人工智能领域的科技论文。这 10 年间，人工智能领域科技论文数量随时间分布如图 2 所：

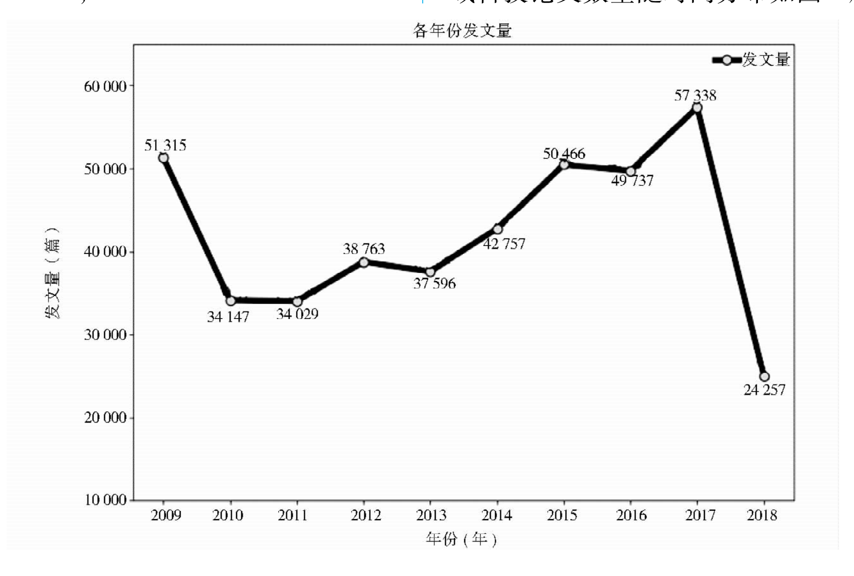


图 2 人工智能领域科技论文的时间分布示意

2.3 数据清洗与规则构建

2.3.1 机构数据的清洗

科研团队识别的前提是对作者数据进行消歧，而作者消歧需要与机构数据结合，因此，我们首先要对机构数据进行清洗，再进行作者数据清洗。在已有研究与实践基础上，制定机构数据清洗的流程与规则如下：

（1）依据机构所属国家名进行区分。提取论文机构和国家名称，若机构名称相同但国家名称不同，则视作不同机构。

（2）采用迭代式积累方法设计清洗规则并进行消歧[10]，具体为：第一，机构名称前后顺序不同，实际上是同一家机构，例如“Washington Univ”和“Univ Washington”是同一家机构；第二，同一机构下属的不同实验室或机构，合并到同一机构，例如“NICTA Canberra Lab”和“NICTA Queensland Lab”同属于机构“NICTA”；第三，国内大学英文名称和拼音名称等不同类型的名称，统一使用该大学的拼音名称，例如“Beijing Univ Aeronaut & Astronaut”统一称为“Beihang Univ”；第四，同一机构既有缩略名，也有全名，则将所有类型的缩略名统一为该机构的全称，例如“EPFL，Switzerland”、“EPFL IC”，清洗后名称为“Ecole Polytech Fed Lausanne，Switzerland”。

（3）人工核查。利用分数计数法计算出各个机构的发文量，并依据发文量进行降序排列，共有机构 17 511个，人工对发文量前 1% 的机构进行核查，合并基于前述规则未能发现的同一机构不同表述的情形，并以发文量较高的机构名称为准。

2.3.2 作者数据的清洗

作者姓名歧义（Author Name Ambiguity）问题在科学计量分析中广泛存在，一般地，这一问题可以细分为“异形同义”和“同形异义”两类。前者指同一个作者拥有多种不同形式的姓名写法，如全称写法和缩略写法；后者指普遍存在的重名现象，这一问题在亚洲学者中尤其严重。为保障数据分析的质量，在机构清洗的基础上，制定以下数据清洗流程和规则来实现作者人名消歧[11]，具体如下：

（1）数据格式转换。将 WoS 格式的原始数据转换为可供后续处理的格式。

（2）抽取待消歧作者机构与合著者信息。这里使用前面已经清洗好的机构信息。

（3）依据机构信息和合著者信息实现作者消歧。第一，机构相同的重名作者视为同一作者；第二，不同机构之间的重名作者若存在相同合著者，则判定为同一作者实体；第三，不满足上述两个条件的重名作者判定为不同作者，用“下划线 + 数字”的方式进行区分。

未消歧之前有 53 万名作者，由于重名作者的存在，消歧之后得到 65 万名作者，最终结果存储在结构化表单中。

2.4 科研团队识别

2.4.1 整体合著网络的构建

在进行作者姓名消歧之后，我们构建了数据集中所有合著文章的整体合著网络，共涉及656 668 个节点(作者)和2 042 924 条边，边的权重代表合作强度，由二者的总合作次数计算得到。相关研究表明，分数计数法比一般使用的全计数法更能发现网络中的簇群[12]，也与实际科研产出更相关[13]。因此，本研究使用分数计数法来计算合作次数，即若一篇文章有 n 个作者，则任意两个作者之间的合作次数为 1/n。由于构建出的合著网络规模巨大，因此无法进行可视化展示，后续所有操作都是利用程序完成。

针对初始网络，我们采用 Pajek 中 Louvain 方法。选择"Multi-Level Coarsening + Multi-Level Refinement"，其他参数为默认值，共探测到 94 347 个社区。这些社区紧密相联，又与其他社区相分离，因此视作以合著关系形成的科研团队。对这些科研团队进行描述性统计分析，发现最大的团队包含 1 553 位作者，规模前 10 的团队中，仅有一个在 1 000 人以下(996 人)。这样的结果无法满足我们的细粒度分析需求，并且实际的科研团队规模与我们的识别结果有所出入，这可能是因为科研团队当中存在大量单次合著、重要性较低的边缘结点，它们被误认为是团队成员。因此，我们需要对原始合著网络进行提取。

在本研究中，我们选择删掉原合著网络中发文量为 1 并且被引量低于 100 的作者结点，因为这些作者在人工智能领域研究的影响力非常弱，甚至并不是真正属于人工智能领域的学者，例如，是参与部分非核心工作的研究生或技术人员，在科研团队中处于极其边缘的位置，因而构成科研团队识别的干扰因素，需要将其剔除。提取后，得到节点数为 186 997、连接数为 543 351的新合著网络，与原始合著网络相比，节点数减少了 469 671，连接数减少了 1 499 573，整体合著网络在保持原有重要节点的基础上大幅度缩小。

2.4.2 识别粒度的选择

为了识别出规模合理、可供分析的科研团队，需要不断地调整参数，对不同参数下所识别出的科研团队结果进行对比和评估。由于合著网络规模巨大，每一次参数调整重新计算都需要耗费大量时长的计算资源。随着参数的调整，识别出的科研团队逐渐趋于稳定。聚类识别采用的是 Louvain 算法，Resolution 参数影响识别出聚类的规模，Max Level 和 Max Iteration 两个参数对应算法的迭代条件限制。以已知人工智能科研团队作为参照进行调参，选取了西安电子科技大学焦李成科研团队，调研了其科研团队成员构成。通过调参，当参数设置为 Resolution = 290，Max Level = 13，Max Iteration = 13 时，达到了较为适宜观察团队内部具体情况的粒度，得到的焦李成科研团队与实际调研所得到的科研团队基本相同，此时得到的科研团队规模均在 100 人以内。如果团队规模过大，可能内部联接会较为松散。如果团队规模过小，可能会遗漏掉某些重要联接。需要说明的是，在团队规模粒度选择上并没有严格的标准，只是所揭示的科研团队紧密程度会有所不同。

在进行参数调整的过程中，我们也对团队来源的合理性进行了验证，以确保识别出的团队是来自上一个更大的团队，见图 3。

2.5 领军团队的提取

人工智能领域有很多科研团队，但在实际研究和工作中，主要关注处在领军位置的科研团队。因此，在识别科研团队的基础上，需要进一步提取出领军团队。我们采用 6 种指标从不同角度对科研团队进行测度，分别是发文量(Number of Publications)、被引量(Number of Citations)、h 指数(h index)、加权点度中心度(Weighted Degree Centrality)、中介中心度(Betweenness Centrality)和接近中心度(Closeness Centrality)。其中，前三种指标从结点属性上测度科研团队的实力，后三种指标从网络结构上测度科研团队的实力。发文量、被引量和 h 指数的具体数值根据其定义，采用自编 Python 程序计算获得，加权点度中心度、中介中心度和接近中心度指标的具体数值，利用大型社会网络分析工具 Pajek 计算得出。对于每个维度的指标，取排名前 10 位的科研团队作为该维度的领军团队。

3 研究结果与分析

3.1 科研团队的整体情况

基于上述过程，共识别出人工智能领域科研团队 23 423个，涉及作者 186 997 名，团队规模的分布情况见图 4。许治等[14]认为，小规模团队(10 人以下)在合作网络密度与合作强度上均优于大规模团队(25 人以上)。由图 4 可以看到，团队规模在 25 人以内的团队占比达到 89.4%，其中 10 人以下的团队规模占比为 78%。团队规模分布较为合理。

图 3 经调参团队 1 分裂成粒度更小的团队 2 和团队 3

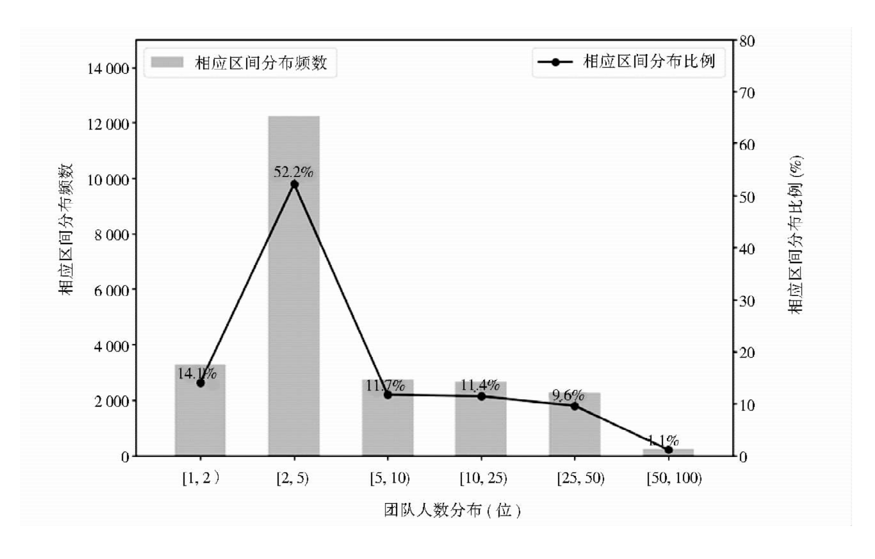


图 4 人工智能领域科研团队按规模分布

余厚强,白宽,邹本涛,等.人工智能领域科研团队识别与领军团队提取[J].图书情报工作,2020,64(20):4-13.

其次,从发文量、被引量、h 指数、中介中心度、接近中心度以及加权中心度六个维度对科研团队进行测度和排行,将前 234 名(占总数前 1%)科研团队定义为领军团队,每个指标分别选取排名前 10 的领军团队作展示和分析,共涉及 47 个团队,因为其中有 10 个团队在两个或两个以上的指标排名中进入前 10,其中团队#205、#342、#207 分别在三个指标的排名中进入前 10。由于篇幅限制,仅对排名前三的科研团队作简明扼要的分析,以及对排名第一的领军团队结构作可视化展示。

3.2 基于发文量的领军团队识别

基于发文量的领军团队见表 1,可以发现发文量高的团队,成员数量基本在 50 人以上。以排名第一编号为#448 的团队为例,该团队共有 52 位有紧密合作关系的著者,见图 5,该团队的领军学者为 Pedrycz Witold _2(末尾数字是对数据分析中对重名作者的编号)。该领军团队的研究前沿主要关注时间序列(time series)、模糊认知图(fuzzy cognitive maps)、数据挖掘(data mining)三个方面。编号为#1927 的团队共有 54 位有紧密合作关系的著者,该团队的领军学者为 Zhang Mengjie_2,研究前沿主要关注边缘检测(edge detection)和机器学习(machine learning)两个方面。编号为#1064 的团队共有 58 位有紧密合作关系的著者,该团队的领军学者为 Castillo Oscar,研究前沿主要关注基于模糊逻辑的动态参数自适应问题。

表 1 基于发文量的领军团队

排名	团队编号	团队人数(位)	发文量(篇)	人均发文量(篇)
1	#448	52	242.9	4.7
2	#1927	54	219.2	4.1
3	#1064	58	213.5	3.7
4	#205	58	211.3	3.6
5	#342	49	207.7	4.2
6	#203	73	205.2	2.8
7	#594	77	198.9	2.6
8	#1800	58	197.7	3.4
9	#170	60	191.9	3.2
10	#3661	52	190.9	3.7

3.3 基于被引量的领军团队识别

基于被引量的领军团队见表 2,可以发现被引量高的团队,成员数量相差较大,就排名前十的团队而言,成员最少的#5997 号团队仅有 23 人,成员最多的#843号团队却有高达 80 人。以排名第一编号为#2096 的团队为例,该团队共有 36 位有紧密合作关系的著者,见图 6,该团队以 Lin Chin-Jen 为核心。该领军团队的研究前沿主要关注与人工智能相关的算法研究,主要体现在分类算法和大规模线性分类两个方面。排名第二的编号为#5330 的团队以学者 Sun Jian_14 为核心,研究前沿主要关注基于图像分类(image classification)新模型或新方法的提出。排名第三的编号为#1959的团队以学者 Ma,Yi_4 为核心,研究前沿主要关注计算机视觉领域,探讨的问题包括在严重损坏的情况下识别图像、在图像中找到固定物件的模型等,并试图用矩阵的方法解决这些问题。

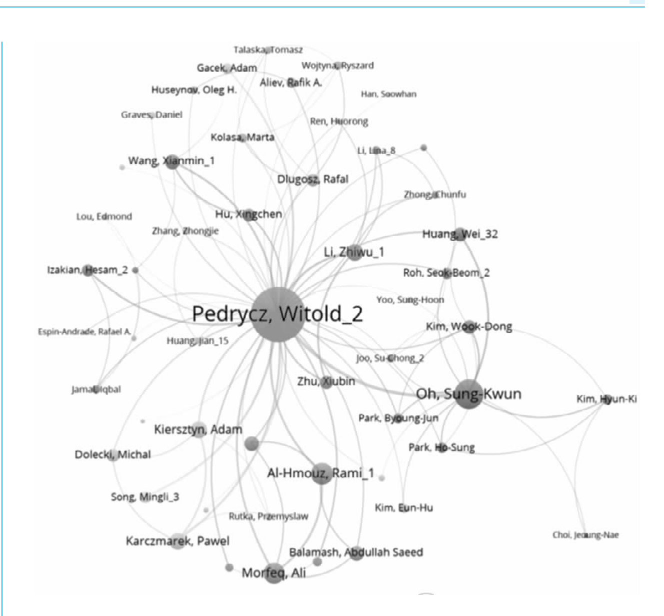


图 5 发文量视角下编号#448 领军团队合著网络

表 2 基于被引量的领军团队

排名	团队编号	团队人数(位)	被引量(次)	人均被引量(次)
1	#2096	36	13 595.7	377.7
2	#5330	36	9 695.1	269.3
3	#1959	25	8 565.5	342.6
4	#929	34	7 887.3	232.0
5	#342	49	7 337.7	149.7
6	#5997	23	6 919.5	300.8
7	#399	43	6 813.1	158.4
8	#843	80	6 775.7	84.7
9	#2435	60	6 664.7	111.1
10	#205	58	6 614.2	114.0

3.4 基于 h 指数的领军团队识别

基于 h 指数的领军团队见表 3,可以发现 h 指数高的团队,成员数量相对较多,最大的#207 号团队有 98 名成员。以排名第一编号为#594 的团队为例,该团队共有 77 位有紧密合作关系的著者,如图 7 所示,该团队以英国布鲁内尔大学 Wang Zidong 为核心,研究前沿主要在同步控制(synchronization control)、多目标优化(many-objective optimization)方面,并且对神经网络在



图 6　被引量视角下编号#2096 领军团队合著网络

时变时滞(time-varying delays)影响下的指数稳定性(exponential stability)进行了探讨。编号为#205 的团队以西班牙格拉纳达大学的 Herrera Francisco 为核心,研究前沿主要关注群体决策(group decision)、基本分类器(base classifier)和分类系统(classification system)三个方面。编号为#108 的团队共有 80 位有紧密合作关系的著者,研究前沿主要关注基于模糊关联规则挖掘和模糊逻辑结合的物流、医疗、仓储等方面。

表 3　基于 h 指数的领军团队

排名	团队编号	团队人数(位)	团队 h 指数
1	#594	77	250
2	#205	58	193
3	#108	80	189
4	#795	85	178
5	#342	49	177
6	#207	98	174
7	#223	51	173
7	#203	73	173
9	#2711	60	171
10	#29	49	169

3.5　基于中介中心度的领军团队识别

基于中介中心度的领军团队见表 4,可以发现中介中心度高的团队,成员数量基本在 40 人以下。以排名第一编号为#698 的团队为例,该团队共有 44 位有紧密合作关系的著者,见图 8,该团队以重庆大学 Zhang, wei_27 为核心,研究前沿主要在图形识别的相关技术,近年研究重心在利用线段匹配的方法提高图像匹配的正确率。编号为#1348 的团队以学者 Willmann,T 为核心,研究前沿关注与人工智能相关的算法研究,特别是

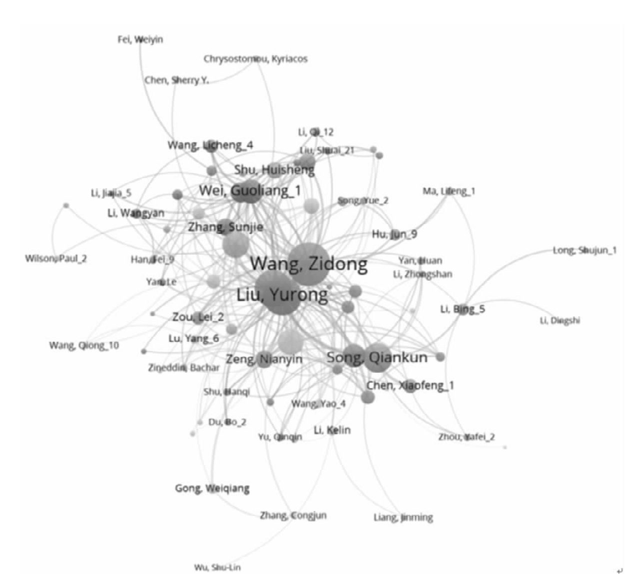


图 7　h 指数视角下编号#594 领军团队合著网络

与矢量量化算法(LVQ)相关的研究。

表 4　基于中介中心度的领军团队

排名	团队编号	团队人数(位)	团队中介中心度
1	#698	44	0.014 8
2	#1348	16	0.010 3
3	#3127	22	0.007 8
4	#904	26	0.005 9
5	#2982	33	0.005 8
6	#499	39	0.005 1
7	#1663	31	0.004 7
8	#2242	38	0.004 5
9	#196	37	0.004 1
10	#63	58	0.003 8

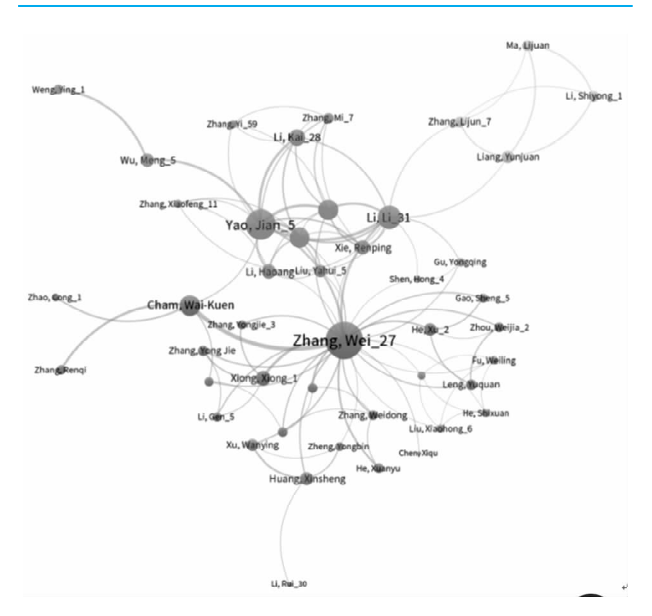


图 8　编号#698 领军团队合著网络

3.6 基于接近中心度的领军团队识别

基于接近中心度的领军团队见表5,可以发现接近中心度高的团队人数普遍较多,成员数量基本在80人以上。以排名第一编号为#2352的团队为例,该团队共有100位有紧密合作关系的著者,见图9,该团队以埃及开罗大学Hassanien,Aboul Ella为核心,研究前沿主要关注支持向量机参数优化的量子粒子群优化,同时还关注贝叶斯优化方法(Bayesian Optimization Approach)、多目标优化算法、求解全局优化问题的可动阻尼波算法等。编号为#2733的团队以美国圣母大学D'Mello, Sidney为核心,研究前沿主要针对教育领域,应用计算机视觉技术(computer vision technique)捕捉学习者在环境下的面部表情并分析其状态,研究的目的是完善他们的智能辅导系统(intelligent tutoring system),因此该团队在语音识别、情绪分析方面颇有建树。编号为#1063的团队以西班牙马德里理工大学Bajo, Javier为核心,研究前沿主要关注物联网系统的自适应容错跟踪控制算法、提高物联网系统的区块链管理效率的非线性自适应闭环控制系统、物联网多设备分布式连续时间故障估计控制。

表5 基于接近中心度的领军团队

排名	团队编号	团队人数(位)	团队接近中心度
1	#2352	100	4.723
2	#2733	83	4.395
3	#1063	84	4.082
4	#2181	94	4.036
5	#795	85	4.001
6	#207	98	3.974
7	#948	84	3.962
8	#5899	79	3.962
9	#2177	92	3.909
10	#3481	83	3.895

3.7 基于加权点度中心度的领军团队识别

基于加权点度中心度的领军团队见表6,可以发现基于加权度的团队排名,在团队人数上分布较为均匀。以排名第一编号为#3127的团队为例,该团队共有22位有紧密合作关系的著者,见图10,该团队以法国巴黎第六大学学者Perny, Patrice为核心,研究前沿主要关注包括公共设施的位置布置问题、电力市场贸易谈判问题、多边谈判的策略问题、碳排放量评估与交易问题、多准则决策问题,核心在于决策理论在人工智能中的应用。编号为#2056的团队以Xu,Yang_7为核心,研究前沿主要是格值逻辑(lattice-valued logic)、格蕴含代数(lattice implication algebra)、SAT问题以及模糊逻辑(fuzzy logic)。编号为#2242的团队共有38位有紧密合作关系的著者,该团队以学者Ramirez,J.和Gorriz,J.为核心,研究前沿主要的是模式识别(Pattern recognition)等人工智能(Artificial intelligence)方法在生物(Biology)和医疗(Medicine)领域的应用研究,近年来特别针对阿尔茨海默症展开研究。

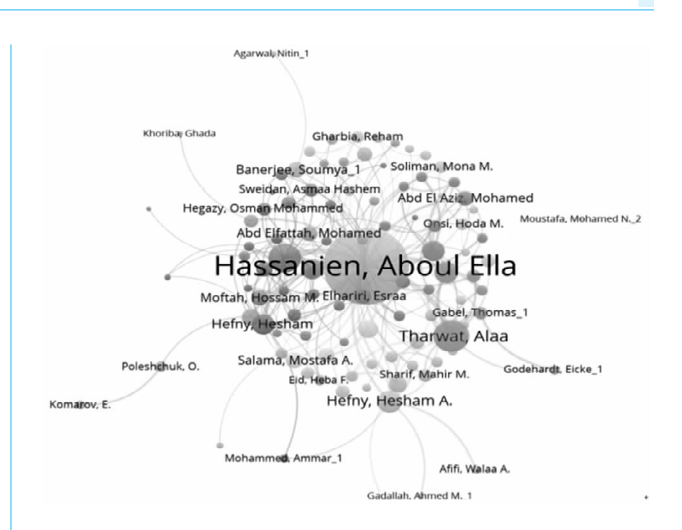


图9 编号#2352领军团队合著网络

表6 基于加权点度中心度的领军团队

排名	团队编号	团队人数(位)	团队加权度
1	#3127	22	423.7
2	#2056	58	413.0
3	#2242	38	399.7
4	#2312	37	393.9
5	#2373	50	393.2
6	#196	37	374.2
7	#207	98	366.6
8	#48	36	355.8
9	#276	61	345.7
10	#2733	83	341.6

3.8 6种维度下领军团队的比较分析

对上述6个维度出发所识别到的领军团队进行比较,结果见表7。表7中,团队#205和团队#342在发文量、被引量和h指数3个研究维度都排到了前10位,团队#207同时在h指数、接近中心度和加权点度中心度都排到了前10位。此外,团队#196、团队#203、团队#2242、团队#2733、团队#3127、团队#594和团队#795均在两个维度中排到前10位。这个结果说明,不同维度的领军团队排名确实揭示了不同内涵的领先优势,与此同时,存在一些领军团队在不同维度均体现出了领先优势,其中,在h指数方面表现优秀的领军团队更可能在其他维度也取得优势。

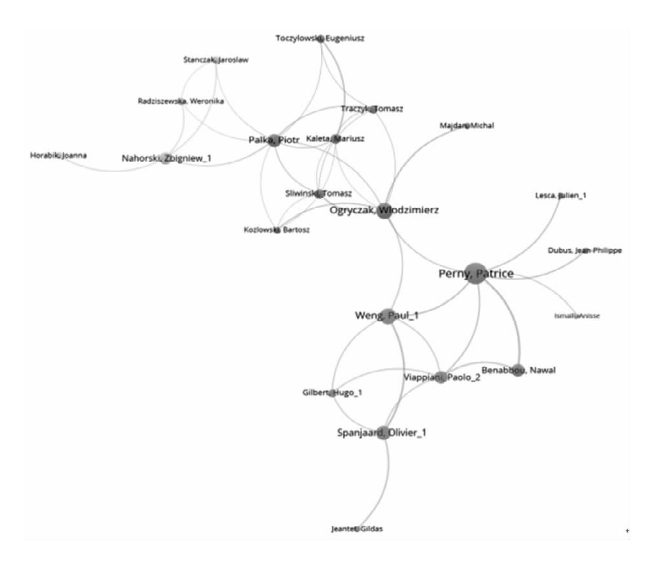


图 10 编号#3127 领军团队合著网络

表 7 6 个维度下领军团队的比较分析

排名	发文量	被引量	h 指数	中介中心度	接近中心度	加权点度中心度
1	#448	#2096	#594	#698	#2352	#3127
2	#1927	#5330	#205	#1348	#2733	#2056
3	#1064	#1959	#108	#3127	#1063	#2242
4	#205	#929	#795	#904	#2181	#2312
5	#342	#342	#342	#2982	#795	#2373
6	#203	#5997	#207	#499	#207	#196
7	#594	#399	#223	#1663	#948	#207
8	#1800	#843	#203	#2242	#5899	#48
9	#170	#2435	#2711	#196	#2177	#276
10	#3661	#205	#29	#63	#3481	#2733

4 研究结论

本文以 2009－2018 年 Web of Science 人工智能学科所有科技论文的数据为来源，构建基于数据分析的从数据清洗、网络构建、科研团队识别与领军团队提取的完整流程。

通过研究，本文主要有如下 3 个主要贡献：

(1)基于迭代式积累设计科技论文数据清洗的规则。形成一套人工智能研究机构别名对应表，提出基于机构名和合著者对作者进行大规模消歧的方法，并在人工智能科技论文数据集上得到实证检验。这种名称消歧的思路和方法较为简便可行，实际消歧效果较好，可以用于其他学科领域的机构或作者名称消歧。

(2)构建基于合著网络关系识别人工智能科研团队的流程体系。采用分数计数法构建全局合著网络，通过消除边缘结点进行合著网络提取，利用已知团队作参考进行动态参数调整，识别出了粒度合适的科研团队。在实际调参过程中，由于没有客观标准，需要选取已知团队作为参照点，判断合适的团队划分标准，这同样适用于其他学科领域科研团队的划分过程。

(3)从 6 个维度提取人工智能研究的领军团队。分别从发文量、被引量、h 指数、中介中心度、接近中心度、加权点度中心度识别了领军团队，并举例分析了领军团队的构成及其研究主题。在本研究中，领军团队用于描述科研成绩突出的科研团队，而科研成绩突出可以体现在不同维度，采用的六个维度指标在科学评价中认可度较高，可用于其他学科领域的科研团队的分析。但是，除了这些维度的指标之外，还有其他维度的指标，可以在未来进一步探索。

参考文献：

[1] 陈春花，杨映珊. 基于团队运作模式的科研管理研究[J]. 科技进步与对策，2002(4)：79－81.

[2] ACEDO F J, BARROSO C, CASANUEVA C, et al. Co-authorship in management and organizational studies: an empirical and network analysis [J]. Journal of management studies, 2006, 43(5): 957－983.

[3] GREGORIO G, JINSEO P, CHARLES H, et al. Scientific author-

ships and collaboration network analysis on chagas disease: papers indexed in pubmed (1940-2009) [J]. Journal of the institute of tropical medicine in sao paulo, 2012, 54(4): 219-228.

[4] 李亮,朱庆华. 社会网络分析方法在合著分析中的实证研究[J]. 情报科学,2008, 26(4):549-555.

[5] 李纲,李春雅,李翔. 基于社会网络分析的科研团队发现研究[J]. 图书情报工作,2014,58(7):63-70,82.

[6] 沈耕宇,黄水清,王东波. 以作者合作共现为源数据的科研团队发掘方法研究[J]. 数据分析与知识发现, 2013,29(1):57-62.

[7] 吕璐成,赵亚娟,王学昭,等. 基于关联规则挖掘的研发团队识别方法 [J]. 科技管理研究, 2016, 36(17):148-152,189.

[8] ANTONIO P, CARLOS O, FÉLIX M. Detecting, identifying and visualizing research groups in co-authorship networks[J]. Scientometrics, 2010, 82(2):307-319.

[9] 任妮,周建农. 合著网络加权模式下科研团队的发现与评价研究[J]. 现代图书情报技术, 2015,31(9):68-75.

[10] 范丽鹏,余厚强,姜宇星,等. 人工智能研究前沿识别与分析:基于高产机构对比研究视角[J]. 情报理论与实践,2019,42(9):16-21.

[11] TRAN H N, HUYNH T, DO T. Author name disambiguation by using deep neural network[C]//Asian conference on intelligent information and database systems. Phuket, Thailand: ACIIDS, 2014: 123-132.

[12] GLANZEL W. National characteristics in international scientific co-authorship relations[J]. Scientometrics, 2001, 51(1):69-115.

[13] PRITYCHENKO B. Fractional authorship in nuclear physics[J]. Scientometrics, 2016, 106(1):461-468.

[14] 许治,陈丽玉,王思卉. 高校科研团队合作程度影响因素研究[J]. 科研管理,2015,36(5):149-161.

作者贡献说明:

余厚强:收集数据,提出研究思路,设计研究方案,论文修改;

白宽:处理数据,论文撰写;

邹本涛:数据获取,数据分析,论文修改;

王曰芬:论文整体布局,提出修改意见,论文定稿。

Identification and Extraction of Research Team in the Artificial Intelligence Field

Yu Houqiang Bai Kuan Zou Bentao Wang Yuefen

School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094

Abstract: [**Purpose/significance**] This paper identifies the research team in the artificial intelligence field, and extracts the leading research team from multi-dimensional indicators, aiming to enrich the process and method of identification of the research team, and provide the basis for analyzing the context, frontier and theme of the field of artificial intelligence from the perspective of the research team. [**Method/process**] This paper was based on the publication data of the Web of Science category *Computer Science*, *Artificial Intelligence* from 2009 to 2018, and did data cleaning via programming and manual check. Global co-author network is constructed based on the fractional counting method, and the Louvain algorithm was used to dynamically tune and identify the research teams. Moreover, the leading research team was extracted based on different indicators with parameter adjustment. [**Result/conclusion**] From practical view, the study has constructed a set of rules for cleaning publication data of artificial intelligence field. The process of identifying artificial intelligence research teams based on co-authorship is constructed. The study proposes the method of tuning the parameter by eliminating edge nodes in the collaboration network and further taking the known research teams as baseline. The worldwide research teams of artificial intelligence field are systematically and accurately identified. The leading research teams are further extracted based on indicators of six dimensions, i. e. number of publications、number of citations、h index、weighted degree centrality、betweenness centrality、closeness centrality. Exemplary analysis is conducted on leading research teams of each dimension by combining the publication data and web information survey.

Keywords: artificial intelligence co-authorship network research team leading research team data analysis